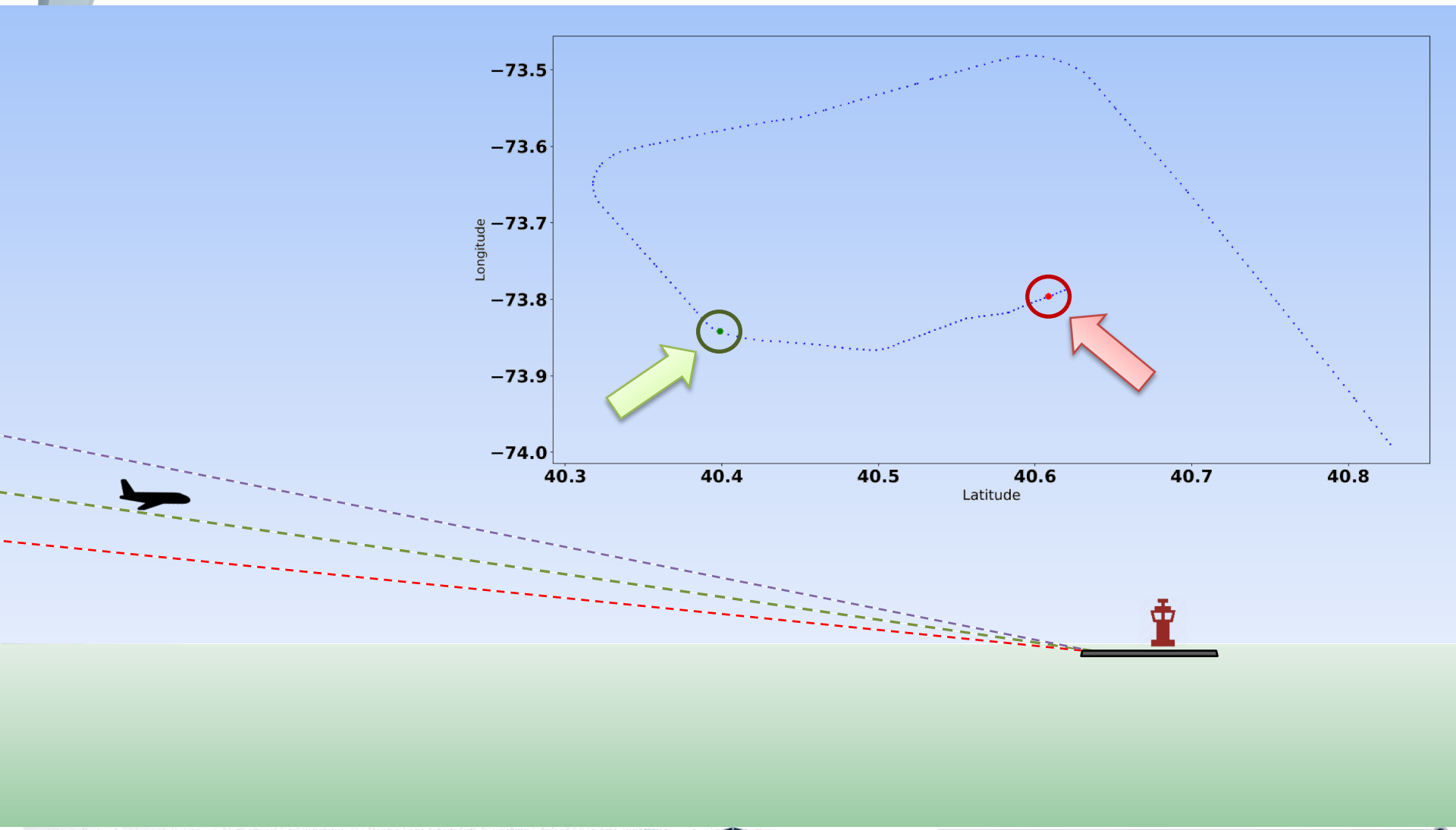
EXPLAINABLE MACHINE LEARNING FOR AVIATION SAFETY ASSURANCE

November 29, 2018

Intelligent Automation, Inc.	**Alexander Grushin** **Jyotirmaya Nanda** **Ankit Tyagi** **David Miller** **Joshua Gluck**
NASA Ames	**Nikunj Oza**

Degraded States in Aviation



Finding Precursors to Degraded States

1. Detect degraded states in aviation data
   - Such states may increase the likelihood of a safety incident
   - We use statistical methods, potentially with a human in the loop

2. Predict that a degraded state may occur in the future
   - If the prediction is made early, then corrections can be made
   - We train a “black box” recurrent neural network to make the prediction

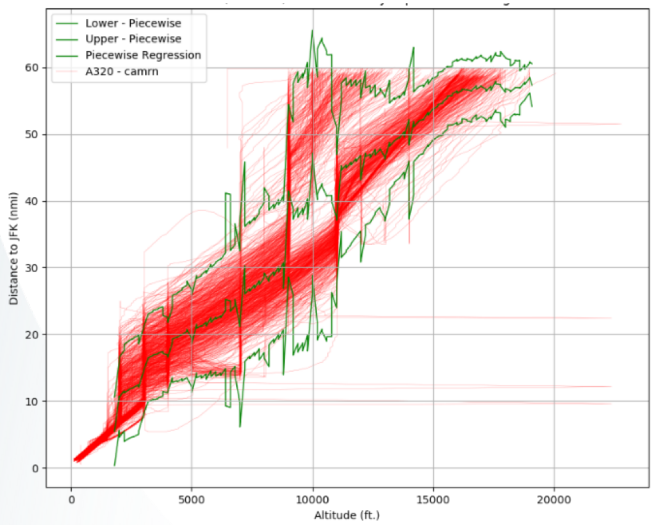
3. Explain why the prediction was made
   - This helps to identify the precursors to the degraded state
   - We extract an interpretable (“white box”) model from the neural network

Data

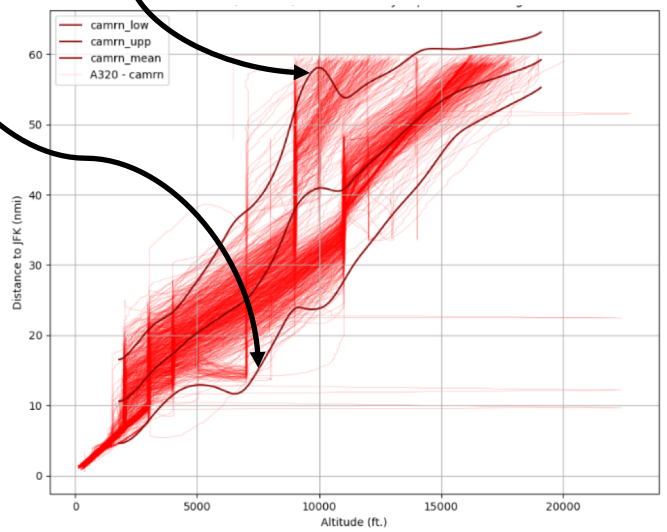
- Sequence $X = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})$ of aircraft state observations over time
- Each observation x_{t_i} contains the multiple feature values:
  - Absolute distance to the airport
  - Relative distance to the airport
  - Altitude
  - Ground speed
  - Latitude
  - Longitude
  - Vertical speed
  - Ground acceleration
  - Heading
  - Heading rate

Detection via Statistical Techniques

- For a given altitude, is the aircraft too far away from the airport? I.e., is relative distance to the airport $\frac{d_{t_i} - l_{t_i}}{u_{t_i} - l_{t_i}} \geq 1$?
  - d_{t_i} is the absolute distance to the airport
  - u_{t_i} is the distance upper bound
  - l_{t_i} is the distance lower bound



Piecewise Linear Regression

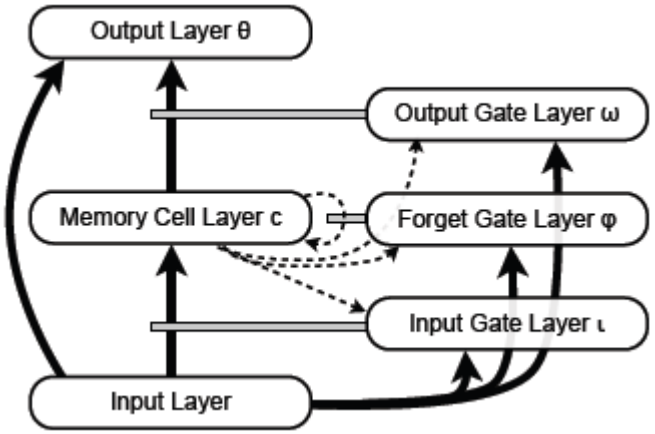


Smoothing via Multiquadric RBFs

Prediction via Long Short-Term Memory

- Input: observation x_{t_i}
- Output: degree of belief $\omega_{t_i} \in [0, 1]$ that a degraded state will occur in the future

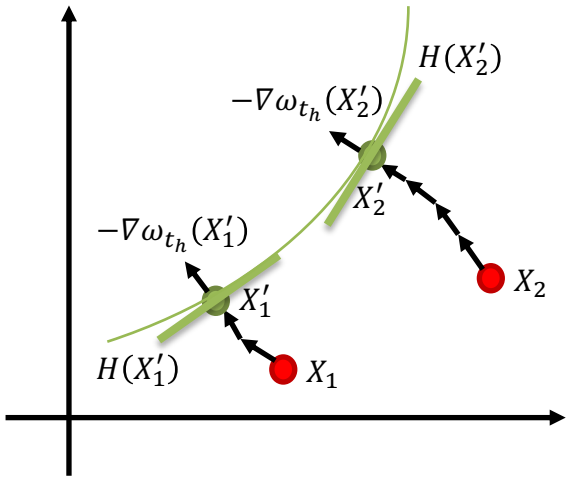
- Memory cells store information for extended periods of time
- Gates determine:
  - How much is stored in memory
  - How long memory persists
  - How memory affects the output



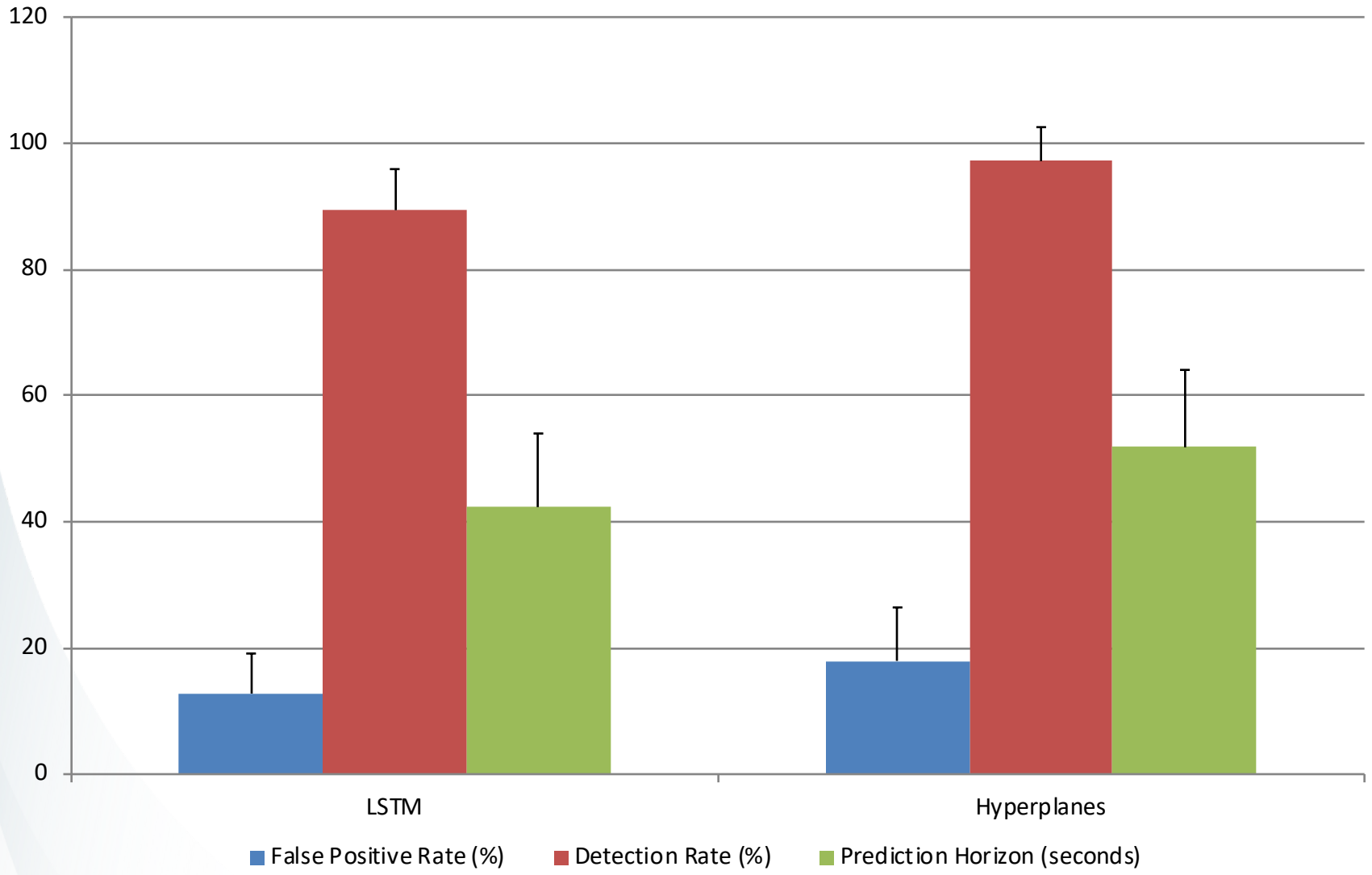
[Monner and Reggia 2010]

Explanation via White Box Model Extraction

- Start with a degraded training sequence $X = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})$
  - The degraded state is predicted at time $t_h \leq t_n$
- Perturb X via gradient descent, until the network's prediction changes
- Let X' be the perturbed sequence
  - Note that X' exists on (or near) the network's decision boundary
- Compute the gradient $\nabla \omega_{t_h}(X')$ with respect to X'
- Define the hyperplane equation: $\nabla \omega_{t_h}(X') \cdot X - \nabla \omega_{t_h}(X') \cdot X' = 0$



Approximation Accuracy



Interpreting the Decision Criteria

Feature	Average Coefficient	Standard Deviation of Coefficients
Distance	5.66	2.49
Relative Distance	14.48	6.12
Altitude	-1.06	0.66
Ground Speed	-7.73	3.39
Latitude	-0.96	0.83
Longitude	0.45	0.78
Vertical Speed	-1.26	0.57
Acceleration	-0.64	0.84
Heading	-0.06	0.26
Heading Rate	0.32	0.33

A distance upper bound violation is more likely to occur if the flight is already close to the upper bound; this may be associated with lower ground speeds

Conclusions and Future Work

- We performed sensitivity analysis at the decision boundary, and approximated it via hyperplanes
  - The approximation is accurate, for the given prediction problem
  - The approximation yields insight into the network's decision-making logic
- We are applying the approach to find precursors for other types of degraded states
  - E.g., unstable approaches
- Can the approach work for other neural network architectures?
  - E.g., deep neural networks

References

1. W. Samek, T. Wiegand and K.-R. Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, *arXiv:1708.08296v1*, 2017
2. D. Monner and J. Reggia, Generalized LSTM-like Training Algorithm for Second Order Recurrent Neural Networks, *Neural Networks* 25(1):70-83, 2012
3. L. Arras, G. Montavon, K.-R. Müller and W. Samek, Explaining Recurrent Neural Network Predictions in Sentiment Analysis, *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159-168, 2017
4. S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks, *CVPR*, pp. 2574–2582, 2016

Acknowledgments

- Intelligent Automation, Inc.
  - Naresh Cuntoor
  - Chris Kurcz
  - Yaroslav Ryabov
  - Bryan Stewart
  - Sergey Voronin
  - Frederick Wieland
- Metron Aviation
  - Jim Smith
- Purdue University
  - Apoorv Maheshwari

Thank You!